Some consideration on expressive audiovisual speech corpus acquisition using a multimodal platform

Sara Dahmani, Vincent Colotte, Slim Ouni

Language Resources and Evaluation manuscript No.
(will be inserted by the editor)

Some consideration on expressive audiovisual speech corpus acquisition using a multimodal platform

Sara Dahmani · Vincent Colotte · Slim Ouni

Received: date / Accepted: date

Abstract In this paper, we present a multimodal acquisition setup that combines different motion-capture systems. This system is mainly aimed for recording expressive audiovisual corpus in the context of audiovisual speech synthesis. When dealing with speech recording, the standard optical motion-capture systems fail in tracking the articulators finely, especially the inner mouth region, due to the disappearing of certain markers during the articulation. Also, some systems have limited frame rates and are not suitable for smooth speech tracking. In this work, we demonstrate how those limitations can be overcome by creating a heterogeneous system taking advantage of different tracking systems. In the scope of this work, we recorded a prototypical corpus using our combined system for a single subject. This corpus was used to validate our multimodal data acquisition protocol and to assess the quality of the expressiveness before recording a large corpus.

S. Dahmani, V. Colotte, **S. Ouni - Slim.Ouni@loria.fr (Corresponding author)**

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

E-mail: FirstName.LastName@loria.fr

We conducted two evaluations of the recorded data, the first one concerns the production aspect of speech and the second one focuses on the speech perception aspect (both evaluations concern visual and acoustic modalities). Production analysis allowed us to identify characteristics specific to each expressive context. This analysis showed that the expressive content of the recorded data is globally in line with what is commonly expected in the literature. The perceptual evaluation, conducted as a human emotion recognition task using different types of stimulus, confirmed that the different recorded emotions were well perceived.

Keywords expressive audiovisual speech · facial expressions · acted speech

1 Introduction

When dealing with expressive audiovisual speech synthesis, acquiring a corpus is an essential step. The corpus textual content should be phonetically rich to cover different diphones in different contexts (previous and following diphones) as recommended in acoustic speech synthesis literature (François and Boëffard, 2001; Volker Strom and King, 2006; Jonathan Chevelu and Delhay, 2008; Dutoit, 2008). Moreover, in the case of expressive speech synthesis, the corpus should cover different emotions. More than that, in comparison with the corpus for acoustic-only speech synthesis, dealing with the visual component of speech is time-consuming, which may constrain the size of the corpus to acquire.

In this paper, we address the issues that we experienced while recording an expressive audiovisual speech corpus. Some information on corpus recording setups and statistics on their content can be found in the literature, but very little information can be found about some essential details for building an audiovi-

sual speech corpora efficiently. We have noticed that the majority of the existing descriptions are related to audiovisual corpora made for audiovisual speech recognition (a list of several corpora can be found in (Fernandez-Lopez and Sukno, 2018; Czyzewski et al, 2017)), very few describing corpora for audiovisual speech synthesis (Barra Chicote et al, 2008; Schabus and Pucher, 2014; Ouni et al, 2013; Mattheyses et al, 2009) and less using 3D tracking technologies (Busso et al, 2008; Kawaler and Czyzewski, 2019). As the main goal is creating a system for expressive audiovisual speech synthesis, the corpus should contain well recognizable emotions and should capture acoustic and visual speech articulation accurately.

To acquire facial data related to speech, optical motion-capture technique can be used (Ma et al, 2006; Vatikiotis-Bateson et al, 1993; Jiang et al, 2002; Kawaler and Czyzewski, 2019) for its good accuracy and appropriateness for kinematic analyses of lip/jaw movements (Feng and Max, 2014; Ouni and Dahmani, 2016). Although this technique is widely used for tracking facial expressions, it can, however, track only facial parts where the markers are visible to the cameras and it cannot track finely the inner mouth region. Markers located on the lips, which are occluded during protrusion or mouth closure, can disappear or be erroneously matched with the wrong side (lower or upper) of the lip. This remark holds for markerless techniques, like kinect-like systems. This type of system is based on face-tracking algorithms that are mostly limited by their low frame rates, their lack of precision in depth and are not suitable for smooth speech tracking (Ouni et al, 2013; Bandini et al, 2015). Moreover, face tracking algorithms are based on approximation and not on real marker spatial position (Keselman et al, 2017). For all these reasons, we decided to use an optical motion-capture system, Vicon, to track the cheeks, the nose, and the forehead regions.

To track the lips accurately, we choose an Electromagnetic articulography (EMA) system. EMA is a prominent method to record speech-related articulation, it exploits the physical properties of electromagnetic induction to track the position of the articulators in three dimensions (3D) with a high temporal resolution. Such a system allows us to track tiny sensors attached to speech articulators such as the tongue, teeth, and lips (Katz et al, 2014; Mefferd, 2015; Walsh and Smith, 2012). This technique is widely used by the speech community and different articulographs were recognized as accurate systems (Stella et al, 2013; Berry, 2011; Yunusova et al, 2009). Yet, EMA system comes with a limitation: the maximum number of sensors is 24 only. This constraint can be overcome by using EMA system to track the movement of the lips only. In fact, lip deformation can be classified into three action components: opening-closing, narrowing-spreading, and protrusion non-protrusion. The latter is the most complex gesture that is difficult to track using only visible markers. The hidden inner lip shape influences the outer appearance. Thus, we use EMA to track visible and partially visible parts of the lips. The other facial parts will be covered by different tracking systems.

Finally, and since all these acquisition techniques are intrusive and based on markers/sensors, a markerless system was used to track the shape of the actor eyes. For this task we choose the RealSense camera with Intel SDK for face tracking. The combination of all those systems in one unique acquisition platform was not a straightforward task. The process of combination will be detailed in this article.

After defining the textual content of the corpus, the way it should be uttered and recorded, and before recording a big corpus, it is important to ensure that the expressed emotions are well perceived. This step is crucial to assess the quality of the expressiveness of the corpus itself, before tackling the synthesis process.

Actually, a not sufficiently expressive corpus or containing wrong expressions can be very harmful to the result of the synthesis. For this reason, we suggest recording a small corpus covering the different emotions and then perform an evaluation of the acquired data. The evaluation is related to: production and perception. The production analysis provides more information on the characteristics specific to each expressive context. Perceptual evaluation helps us to assess the quality of the expressiveness of the corpus and how it is perceived during a human recognition task.

To sum up, the aim of this paper is to study the feasibility of merging three heterogeneous systems (EMA, VICON and RealSense) to record a small expressive audiovisual speech corpus. This prototypical corpus is analyzed to assess the ability of our system to record emotionally reliable audiovisual data with fine lip tracking. This corpus is a preliminary step before recording a larger corpus for speech synthesis. First, we present our multimodal acquisition system that will be used to acquire the expressive audiovisual speech, and how it will be processed. Then, we present a visual-acoustic analysis of the expressive data, followed by a perceptual evaluation that allowed us to formulate conclusions concerning the emotional content of the recorded corpus and to consider the recording of a larger corpus.

2 Expressive audiovisual speech corpus

2.1 Multimodal Acquisition System

Our multimodal acquisition platform is composed of (1) a motion-capture system (VICON) using optical reflective markers, (2) an articulograph (AG501) using

electromagnetic sensors, and (3) a markerless (without physical markers) motion-capture system (Intel RealSense, a depth camera) that allows us to track the movement of the face without markers. The idea of combining different acquisition systems in one multimodal acquisition platform, came from our will to obtain a high quality 3D audiovisual expressive corpus, by using a well-adapted technique for each part of the face. Capturing the dynamics of facial expressions was of high importance, but obtaining a very accurate articulation movement was also very crucial. In a previous study, we have assessed the precision of these three systems. We have found that AG501 system has the highest temporal and spatial precision followed by the Vicon system and then the RealSense system (Ouni and Dahmani, 2016).

The Vicon system is based on 4 Vicon cameras (MX3+) using modified optics for near range. The cameras were placed at *approx.*,150 cm from the speaker. Vicon Nexus software provides the 3D spatial position of each reflective marker at a sampling rate of 100 Hz. Reflective markers of 3 mm in diameter have been glued on the upper part of the actor's face. They are aimed to capture facial expressions. The articulograph (EMA) sensors allow us to track mainly the lip movement. The EMA technique captures finely these gestures at a sampling rate of 250 Hz. The EMA is known for its great accuracy (Berry, 2011; Stella et al, 2013; Yunusova et al, 2009) and for its ability to track in-mouth articulatory movements such as the tongue but also the inner lip regions especially in the case of labial closing movements (Ouni and Gris, 2018). In fact, handling lips occlusion is not possible with camera-based systems, since the sensors must continually be in the field of view of the cameras (Ouni et al, 2013). The RealSense system was used to capture mainly the shape and movement of the eyes. As this system is markerless, and

Fig. 1 The multimodal platform used to acquire data.

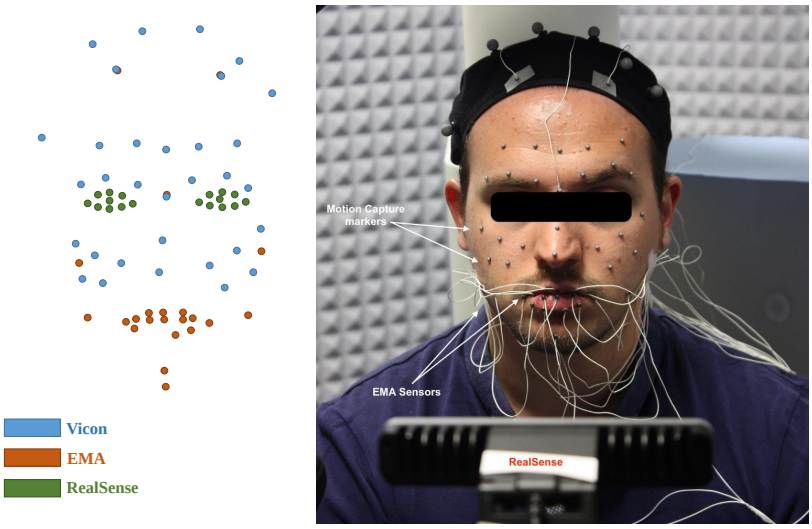

Fig. 2 The positions of the reflective markers, EMA sensors, and RealSense, relatively to the face of the actor and the minimalist face representation obtained after merging the data from the three systems.

that the accuracy of the eyes tracking is not important in the scope of this work, it was appropriate for this task. Its effective sampling rate is 50 Hz.

Figure 1 presents the multimodal platform composed of AG501, Vicon and RealSense systems. Each system comes with different terms of use and placing constraints. The EMA should be placed away from any equipment containing ferromagnetic metallic materials, such as Vicon cameras and tripods, to prevent EM (Electromagnetic) field deformation. In addition, Vicon system should be placed at a reasonable distance from the actor to be able to track the 3mm reflective sensors glued on his face. Furthermore, RealSense depth sensor has a short range (between 20 cm and 120 cm) and thus should be placed closer than Vicon to the actor, with a risk of covering the Vicon cameras field of view. More than that, EMA wires should not hide Vicon markers and should not prevent the RealSense algorithm from recognizing the face shape. Moreover, spatial and temporal synchronizations between the different channels dictate additional combinations and a particular sensors arrangements.

Figure 2 shows the setup where Vicon markers and EMA sensors are placed on the face, and the RealSense is placed in front of the actor. EMA sensors are concentrated around the mouth. The other regions of the face are completed with Vicon sensors. Three Vicon markers and three EMA sensors have been placed in the same positions (each one is on the top of the other, see Figure 3). These three sensors are used as reference sensors to compute the spatial alignment between the two systems. To ensure that the same landmark positions will be saved for the acquisition of the large corpus and for different acquisition sessions, we will print a 3D scan of the actors face with the positions of the markers. We have placed five extra markers on the top of the head to remove the head movement. In fact, involuntary movements influence absolute locations of all markers. We can obtain relative coordinates of these latter, by subtraction of head markers trajectories.

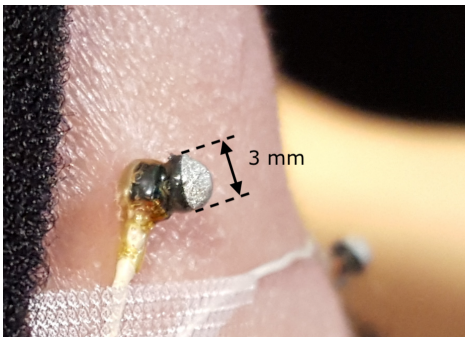

Fig. 3 Vicon marker on the top of an EMA sensor to be used. The marker and the sensor are used as reference points to merge the data (see section 2.5.2).

This way, we keep only the facial movements. Figure 2 presents the layout of the physical and virtual markers, and sensors on the actors face.

The audio was acquired simultaneously with the spatial data using a cardioid microphone (Rode NT3) with 16 bit resolution and 48kHz sampling frequency. To synchronize the audio channel and the motion capture channel, we have used an in-house electronic device. It triggers concurrently an infrared lamp captured by the Vicon system and RealSense, and a high-pitched sound generated by a piezoelectric buzzer for audio captured by the articulograph system (AG501 has an embedded electronic trigger that synchronizes audio and sensor movements). We have developed several tools and algorithmic techniques to process and merge the data accurately, that will be presented in the section 2.5.

2.2 Expressive corpus: Spontaneous or Acted Speech?

The collected corpus of expressive audiovisual speech can be either spontaneous or acted. Spontaneous expressive speech synthesis may reflect a better real situation in oral-based communication. However, recording spontaneous expressive audio-

visual speech corpora may be extremely difficult: (1) finding controlled situations where human speaker expresses spontaneously a given emotion is not easy; (2) the vocabulary and the phonetic coverage cannot be ensured. When dealing with spontaneous expressive speech corpus, it is difficult to control its content (phonetic content and coverage, consistency of the intensity of the emotions, etc.) For these reasons, we propose to use acted speech with different emotions. The purpose of the collected data is not to investigate expressions recognition neither perception (even though characterizing speech in this context is valuable). The virtual talking head is in the same situation as an actor: making an effort to provide convincing and visible expressions, even with some exaggeration. Actually, human expressions are not always visible, and in the majority of cases they are subtle and some human speakers are barely expressive (Hess and Thibault, 2009).

To define acted speech, we may consider collecting dialogues where several expressions can be present. Technically, it is difficult to find sufficient dialogues covering different emotions with various phonetic contexts. In our work, we used a technique called exercise-in-style that was used in the work of Barbulescu (2015).The expression *exercise-in-style* was inspired by the book of Queneau (Raymond, 1947; Queneau, 2018). Using this technique, the actor utters the same sentences while performing six major emotions (joy, surprise, fear, anger, sadness and disgust) and with neutral state. The meaning of the sentences is dissociated from the acted emotions. The actor needs to make an effort to utter each sentence of the corpus with the same expressive intensity, for each emotion. To do this, the actor imagines an emotional situation to be able to express the same given emotion for all the utterances. The main advantage of this technique is that we can still have control of the linguistic content of the corpus. Since we plan to record a big corpus for

speech synthesis in the future, it is not reasonable to construct a specific textual corpus for each emotion. The *exercise-in-style* technique allows us to use the same textual corpus for all the emotions. This way, we can guarantee that the final corpus will be emotionally and phonetically balanced. In this experiment, we choose randomly from a long list of sentences30 sentences (10 short, 10 medium and 10 long). Only the length of the sentences was considered as a selection criteria. For the large corpus, the choice of the sentences will be conditioned by their di-phones (succession of two half phonemes) coverage. In the study of Mehrabian (2008), it was shown that the linguistic content of the sentences contributes only with 7% in communicating the emotional state (the acoustic modality with 38%, and visual modality with 55%). Knowing that, we can suppose that the results of the perceptual tests will reflect the emotion recognition rate even when the linguistic content of the sentences is not completely coherent with the emotional state performed.

To ensure that the quality of the acting will remain the same when recording a larger corpus, we will use the acting technique of Moore (1984), that is based on a list of emotional scenarios for each emotion. The actor has to pick from this list the scenario that feels the closest to his emotional memory. By doing this, we can ensure that: 1) We share the same definition of the emotion with the actor, and later with the participants in the perceptual tests, 2) The actor will have a similar performance of the emotions during the different acquisition sessions.

2.3 Prototypical Corpus

The mini-corpus contains 30 French sentences (10 for short, 10 for medium and 10 for long sentences) for the neutral state and for each one of the six emotions (joy,

surprise, fear, anger, sadness and disgust). The total number of the utterances is 210. The long sentences contain an average of 27 words and lasts around 10 seconds each. The short sentences were 4 to 5 words in length and last about 1.3 seconds.[1]. We recall that the purpose of this mini-corpus is to be used as a prototypical corpus to validate our multimodal data acquisition protocol. Obviously, this corpus cannot be used to develop an audiovisual speech synthesis system, as we probably need more than 20 or 50 times the size of this corpus.

2.4 Acquisition Setup

A 27-year-old, semi-professional, male actor has been asked to utter the different sentences for 7 different emotional states (neutral, joy, surprise, fear, anger, sadness, and disgust). The actor used the *exercise in style* technique. The sentences were presented one at a time on the screen in front of the actor who uttered them showing the same consistent emotion. In this context, the emotions should be considered as acted ones as they are a bit exaggerated as in the case of a play or a movie. Before recording the actor, several preparations have been made. The Vicon system and the EMA system were calibrated. The Vicon markers and the EMA sensors have been glued on the face of the actor. The different systems have been tested before starting the recording. The multimodal acquisition system is thus composed of the Vicon system, the EMA system, the RealSense depth camera and the microphone. We have also placed an RGB video camera to have video reference for the recorded data.

[1] The sentences of medium length have not been used in the different analyses presented in this paper

2.5 Multimodal Data Processing

The motion capture data was obtained directly as 3D coordinates for each marker using a proprietary software that processes the data. Vicon Nexus software provides the 3D spatial position of each reflective marker at a sampling rate of 100 Hz. The AG501 software also provides the 3D spatial position of each sensor at a sampling rate of 250 Hz. The RealSense system has a face tracking procedure, included in the Intel RealSense SDK, that provides the 3D spatial positions of the virtual markers of the contour of the face, the contour of the mouth, the nose, the eyebrows, the contour of the eyes and the pupils, at an effective sampling rate of 50 Hz. Each system provides 3D data on its own reference. For these reasons, it is necessary to define a reference frame and to merge the data from the different systems in this frame. It is also necessary to resolve the problem of temporal synchronization, as each system has its own sampling rate.

2.5.1 Time Synchronization of Multimodal Data

The three systems of 3D data recording come with different maximum sampling rates: EMA (250fps), Vicon (100fps) and RealSense (50fps). Merging the data starts by unifying their frame rates. We choose to conserve the highest frame rate for a better accuracy. We perform an oversampling, using a simple linear interpolation, on the Vicon and RealSense to reach the sampling rate of the EMA (250fps). It has to be noted that the linear interpolation could have been replaced by a more accurate algorithm, as optical flow technique, but this approximation seems sufficient for this work.

To synchronize the different data streams, we use a two-step method. First, we use the information provided by the in-house trigger. The latter generates simultaneously an infrared light-spot and high-pitched sound. We have developed a method that detects the frame where the signal is captured in each stream. For Vicon and RealSense, we look for the first frame where the artificial marker (representing the infrared spot) appears. For AG501, as it has its own trigger, we detect the first acoustic frame where the high-pitched sound appears. This technique offers a reasonable way to synchronize the different streams. However, the result may be off by a few frames, due to the difference in the frame rate of each system. For this reason, we combined this first step by a second to fine-tune the synchronization. During this second step, we synchronize spatially Vicon and EMA data, using the reference markers of EMA and Vicon (glued in the same positions). Our proprietary visualization software allows us to visualize the trajectories of any given marker. We chose a Vicon marker randomly then we can interactively slide its trajectory to match perfectly the corresponding EMA trajectory. At this point we consider EMA, Vicon and the acoustic data synchronized. We repeat this process for RealSense data. As we do not have reference markers for RealSense, we chose a set of virtual markers proposed by the RealSense facial tracking software to associate with eyebrows, the mouth, and the eyes. We visualize again one of the RealSense markers used for the alignment and we try to match its trajectory to that of an EMA sensor. It is important to note that the shifting operation is applied to all the markers of the Vicon or the RealSense even if a single marker was chosen to displace the trajectories. As no drift was detected in the different streams, shifting data in the different channels was sufficient to synchronize them across all the recorded files.

2.5.2 Merging multimodal data

Since each system provides the 3D data on its own reference, the next post-processing step is to define a reference frame and to merge the data from the different systems in this frame. To do this, we used the reference markers of EMA and VICON that we glued in the same place (see Figure 3, for example). Three markers, different from the origin of the reference frame, are needed to construct three non-planar vectors. Then, we calculate the translation and rotation needed to make Vicon data matches EMA data position. These parameters are applied to the entire Vicon sensors and on all the frames of the recording. We did not need to rescale the data, as the three systems provide 3D data on the same scale. Finally, we used extra Vicon markers on the head of the actor and three other EMA sensors (two behind the ears and one between the eyes) to remove head movement. In addition, this kind of data will be used in audiovisual speech synthesis, where removing the head movement is necessary to be able to generate consistent and non-ambiguous facial expressions and speech gestures. It should be noted that it is always possible to reintegrate those movements later on if needed. We use the extra markers to construct the transformation vectors. The first frame of the recording is used as a reference where the head will be fixed at this particular pose. Then we calculate the translation and rotation transformations, based on the configuration of the first frame. After that, we apply the calculated transformation frame by frame. As it can be seen in Figure 2, the final result consists of a 3D point layout representing the face: the mouth, the cheeks, the nose, the contour of the eyes, the pupils of the eyes, the eyebrows, and the forehead. Thus, the consolidated

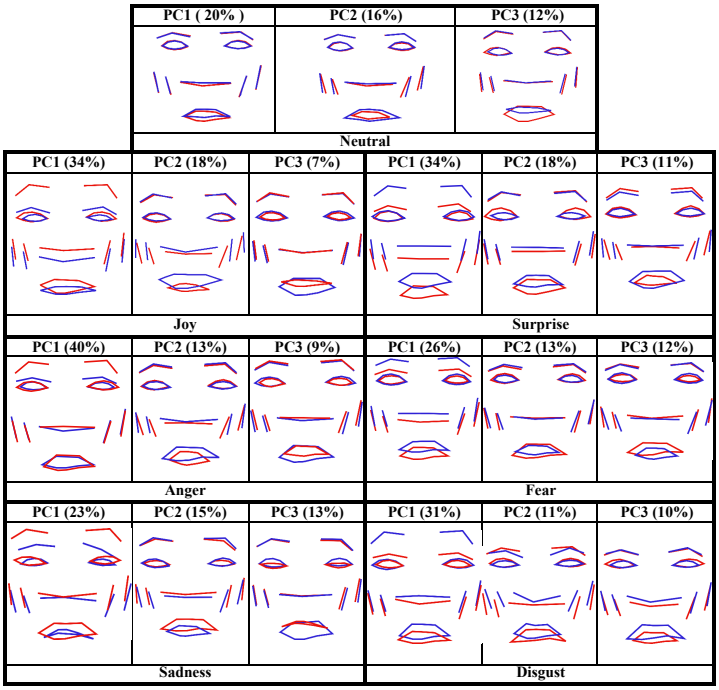

Fig. 4 *The first 3 principal components of the facial data and their percentage of variance for the neutral and the 6 emotions. Each pair of colors shows the deformation of the face when the corresponding component assumes a value of -3 (blue) or +3 (red) standard deviations.*

multimodal data is represented by a 3D point face at a sampling rate of 250Hz, for all the utterances of the mini-corpus.

3 Corpus Expressivity Analysis

When dealing with expressive audiovisual speech synthesis, our goal is to ensure that the expressive speech is correctly perceived by the users and that the acting is convincing. We have analyzed the recorded data to validate the quality of the expressivity produced by the actor. We have conducted two different analyses: (1) an evaluation of the expressive speech production and (2) a perceptive evaluation.

Emotion	Action Units		
	PC1	PC2	PC3
Neutral	AU25	AU18, AU22	AU26
Joy	AU1, AU2, AU5, AU6, AU9, AU12, AU26	AU12, AU18, AU22, AU26	AU26
Surprise	AU1, AU2, AU5, AU26	AU5, AU26	AU18, AU22, AU26
Anger	AU1, AU2, AU5, AU26	AU18, AU22, AU23, AU24, AU26	AU26
Fear	AU1, AU2, AU5, AU26	AU25	AU5, AU18, AU22, AU26
Sadness	AU1, AU4, AU7, AU15, AU25	AU18, AU22	AU7, AU15, AU26
Disgust	AU1, AU2, AU4, AU7, AU10, AU25	AU7, AU9, AU10, AU20, AU25	AU7, AU18, AU22, AU10, AU20, AU25

Table 1 *The first three principal components of the facial data for each emotion and their corresponding Action Units.*

The first evaluation allows us to characterize the expressivity during speech both on visual and audio channels. The perceptive evaluation assesses the quality of the expressivity during a human recognition task, where the purpose is to make sure that this acting is convincing.

3.1 Evaluation of the Expressive Speech Production

In this evaluation, we have used the long sentences only. In a previous study, we have analyzed acoustic and visual data of short sentences in similar conditions (using only the Vicon system (Ouni et al, 2016)). In the following sections, we make reference to the results of that study. In the current work, short and long sentences were only used during the perceptual evaluation. The analyses were made on the obtained data composed of seven files recording (one for each emotion), each

recording contains ten sentences. The measures were made within each recording independently.

3.1.1 Visual data analysis

In the production analysis, we used the same sentences to study the different visual and acoustic emotional features. As the visual data of the recorded corpus consist of 48 spatial points which represent a high-dimensional space, this makes the analysis lengthy and difficult to interpret. Moreover, unlike studying emotions in a static context (images or videos without speech), we are considering the dynamics of the emotions represented by speech sequences. In this particular case, we need to analyze these sequences to extract the most prominent movements present in the recording without analyzing them frame by frame. For these reasons we have performed principal component analysis (PCA) on the data. The corpus has been divided into smaller corpora, where each one represents a set of ten sentences for a given emotion. We have applied the PCA to each corpus to identify the major directions when a given emotion is dominant. In addition, we have also computed several facial distances for each emotion (eyes opening, eyebrows movements, mouth opening and mouth stretching). The goal is to quantify the different variations found with PCA analysis and to represent them with Action Units from the FACS manual (Ekman et al, 2002). The table 2 shows the percentage of variance of the first five principal components of the different emotional states. The variance is distributed over the principal components, but PC1 does not capture a dominant gesture for the different emotions, except for anger. We reach a cumulative percentage of variance greater than 50% with the first 3

Emotion	PC1	PC2	PC3	PC4	PC5
Neutral	20 (20)	16 (36)	12 (49)	8 (57)	6 (63)
Joy	34 (34)	18 (52)	7 (60)	6 (66)	5 (72)
Surprise	34 (34)	18 (53)	11 (64)	6 (71)	5 (76)
Anger	40 (40)	13 (53)	9 (63)	6 (70)	4 (74)
Fear	26 (26)	13 (39)	12 (52)	8 (60)	7 (68)
Sadness	23 (23)	15 (38)	13 (51)	9 (60)	6 (67)
Disgust	31 (31)	11 (42)	10 (53)	8 (61)	8 (70)

Table 2 *Percentages of variance for the first five principal components for the neutral and the 6 emotions. The number in parentheses is the cumulative percentage of variance.*

PCs. In a previous study on short sentences, this threshold was reached with the first 2 PCs only (Ouni et al, 2016).

In Figure 4, we represent the variation of the first three principal components for each emotion. In order to represent the whole variation for each principal component, we represented the minimum and maximum value on each figure. The deformation of the face is presented when the corresponding component has a value of -3 (blue) or +3 (red) standard deviations (we assume they are the lower and upper bound of the component variation). We present in Table 1 the different UAs used for each one of the first three CPs of each emotion. The neutral state represents very small movements, basically lip opening (PC1 with AU25 and PC3 with AU26) and lip protrusion (PC2 with AU18 and AU22). These articulatory movements are important for speech production. This variation concerns only the lower part of the face. The upper part, which is usually important for emotion expression, barely moves. For the first component, and unlike the other emotions where the shape of the eyes varies from wide (AU5) open to slightly closed (AU7),

for *surprise* and *fear* the eyes are wide open continuously. The other noticeable feature is that *sadness* and *disgust* are distinguished by the smallest eye size.

For *sadness*, and in addition to the previous variation, the shape of the eyebrows, the eyes and the mouth also curve downwards. With regard to *disgust*, the mouth is also curved downwards (AU15), but the opening is more important. An important nasal movement is also present in disgust's second component (AU9). 18% of *joy* variance represents an upwards lip stretching characteristic of the familiar smile shape (AU12). It should be noticed that the movement of the face related to speech is also important. Globally for all the emotions, the first three components are related to lip opening (AU25 and AU26) and lip stretching/protrusion (AU18 and AU22). Based on the layout of the reduced representation of the actor's face, some measurements have been calculated and compared based on some specific sensors and distances. In the Figure 5 the Euclidean distance between the sensors (a) and (b) was used to compute the eye opening/closing (AU5, AU7 and AU45). For eyebrow movement, the sensor (d) represents the central sensor of the actor's left eyebrow at rest. The coordinates of this sensor were used as a reference to calculate eyebrow elevation/frowning (AU2 and AU4) using the coordinates of the sensor (c) that represents the central sensor of the actor's left eyebrow for the different emotions. Sensors (e) and (f) were used to measure mouth stretching while sensors (g) and (h) were used to measure mouth opening.

The figure 6 shows the result of the mean eye opening/closing (AU5, AU7 and AU45) for each emotion. This measure is represented in Figure 5 with the distance a-b. The mean value and the standard deviation of each (right and left) eye opening has been calculated. The result of Figure 6 is consistent with our findings using the PCA analysis. The *surprise* and *fear* are the emotions with the highest eye

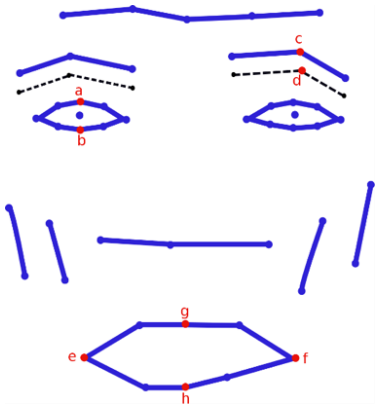

Fig. 5 The layout of the markers used for computing facial measures.

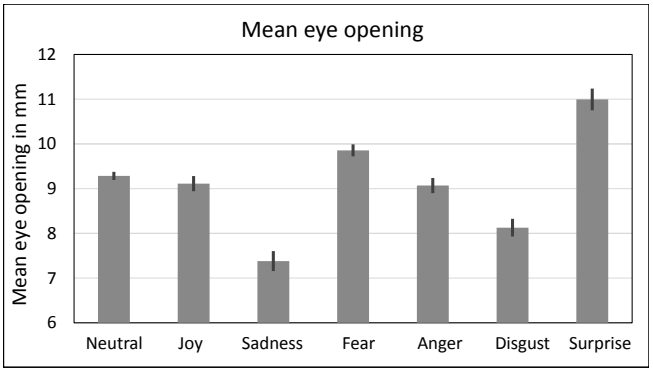

Fig. 6 The mean eye opening (in mm) for each emotion and their standard deviation. This measure is represented in figure 3 with the distance a-b. The mean value and the standard deviation of each eye (right and left) opening has been calculated, then we computed their mean value.

opening value, *sadness* and *disgust* are the lowest and the remaining emotions have a moderate eye opening value.

Figures 9 and 10 show the mean lip stretching and opening for each emotion respectively. The lip stretching is related to the Action Units AU12, AU13, AU14, AU15 and AU20 and has been calculated based on the distance e-f and the lip opening is related to Action Units AU25, AU26 and AU27 and is based on the dis-

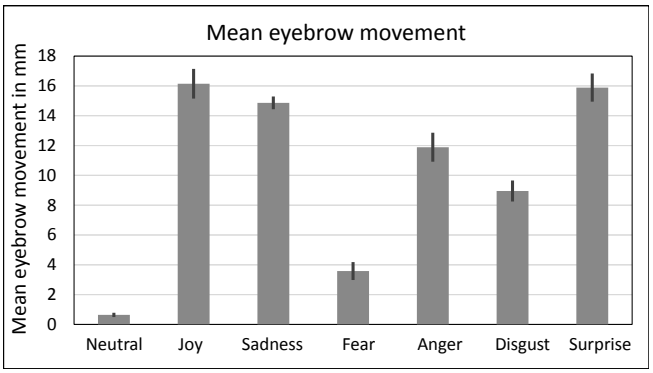

Fig. 7 Mean eyebrow movement in mm for the 7 emotions and their standard deviation. Calculated based on the central eyebrow point (c). A frame at rest position has been selected as reference.

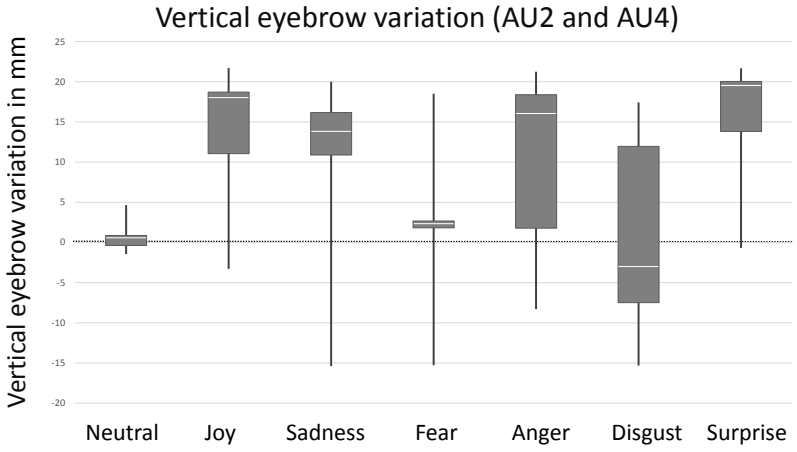

Fig. 8 Vertical axis values in millimeters for the central left eyebrow sensor. The rectangles represent the first and the third quartiles. The horizontal white line represents the median and the extremities of the vertical lines represent the min and max values of the sensor's position. Positive values represent eyebrows elevation (AU2), negative ones represent frowning (AU4).

tance g-h. For the eyebrow movement (Figure 7) is related to Actions Units AU1, AU2 and AU4. This measure has been calculated based on the central eyebrow point (c). A neutral frame has been selected to be used as a reference. This frame represents the position of the eyebrows at rest. The distance c-d has been calcu-

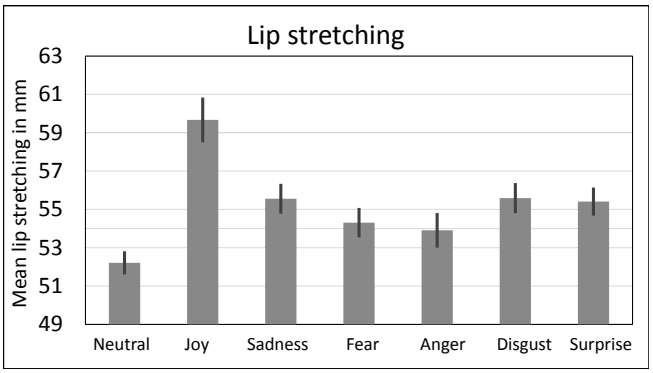

Fig. 9 The average lip stretching (in mm) for each emotion and their standard deviation. The lip stretching has been calculated based on the distance e-f.

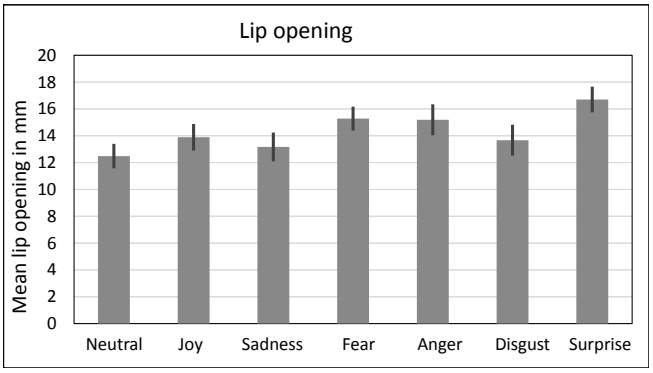

Fig. 10 The average lip opening AU25/AU26, in mm for each emotion and their standard deviation. The lip opening has been calculated based on the distance g-h.

lated for the two eyebrows during each emotion's recording. The mean Euclidean distance c-d has been computed for the two eyebrows, then the mean value of the two eyebrows results has been calculated.

Figures 6 and 10 show that the emotions with the highest mean eyes opening also have a great mouth opening value and inversely. The global tendencies of eyes and lip opening are very similar. As shown in Figure 8, practically for all the emotions, the eyebrows movements concern only elevations (AU1 and AU2),

except the disgust emotion that showed frowning actions (AU4). Some emotions are distinguished by their visual features while others share similar characteristics:

Neutral: The utterances of the emotion were used as a reference to compare all the other emotions features. For this emotion, eyebrows barely move and the other measures have the lowest or very low values.

Surprise: this emotion is distinguished on several sides from other emotions. The eyes and lips opening are the widest (AU5 and AU26). This emotion has also an important lip stretching compared to neutral. Furthermore, the eyebrows values are the highest (AU1 and AU2). These remarks are coherent with the result of PCA analysis, since 34% of the variation represents a continually wide open mouth and raised eyebrows.

Joy: This presentation has the highest lip stretching (AU12). This result can be explained by the smile shape that requires an important lip stretching movement, this shape is clearly captured by the second principal component of *joy*. Nose wrinkler (AU9) is also present in PC1 and PC2. The *joy* presentation has also a great eyebrow movement (AU1 and AU2). For lips and eyes opening (AU25, AU26 and AU5) the obtained values are relatively moderate.

Sadness and disgust: When examining *sadness*, some resemblance to *disgust* emotion was found. They both have the lowest eyes opening (AU7), and an important lip stretching relative to the Action Unit AU15 and AU20 (in consistence with the PCA results, where these two emotions were characterized by downwards lips stretching movement). They have are characterized by the wrinkler movement (AU9) strongly present in the first three principal components. Moreover, they

have a moderate lip opening (AU25). For eyebrows movement, *sadness* is characterized by a steady eyebrow raising movement (AU1 and AU2). On the contrary, in *disgust* performance, the movement varies from a rapid rise of the eyebrows to a steady frowning position (AU1, AU2 and AU4).

Anger: A slight eye opening is present in *anger* PC1. PC2 concern mainly lip tightener (AU23) and lip protrusion movements (AU18 and AU22), which is coherent with the low stretching value of *anger* in figure Figures 9. In addition, in *anger* utterances, lip opening (AU26) was remarkably high (second after *surprise*). Nose wrinkler (AU9) can be noticed in the PC2 and no dominant pattern was noticed for eyebrows movements. The movement alternates steady neutral position and fast or stable elevations (AU1 and AU2).

Fear: This emotion exhibits the most important eyes opening (AU5) after *surprise*. Similarly to *anger*, and despite having low lip stretching value, those emotions compensate by a high lip opening movement (AU26). Figure 7 and 8 show that eyebrow elevation movements (AU1 and AU2) are present but extremely low. By describing the different emotions visual features in terms of Action Units, we conclude that our findings are similar to what is commonly expected (Ekman and Friesen, 1976; Tian et al, 2001; Wiggers, 1982; Lucey et al, 2010). In addition to replicating previous findings, the present study demonstrated that the main facial characteristic of these emotions are maintained even during the speech activity. Moreover, additional Action Units are present in the first three principal components of all the emotions (AU26, AU26, AU18, AU22). Those Action Unites are related to speech activity (mouth opening, protrusion).

3.1.2 Acoustic Data Analysis

The acoustic data was recorded at the same time as the visual data and the two streams were synchronized. We started the post-processing by making a speech alignment for each sentence at different levels: words, syllables and phonemes. By this alignment, we have sought to obtain acoustic characteristics at a very fine level. We used CMUSphinx (an open source toolkit for speech recognition and alignment by Lamere et al (2003)) to do a first phonetic alignment, then we made a manual check, to correct the eventual errors and imperfections. We used PRAAT software (a free software package for the scientific analysis of speech in phonetics Boersma et al), to calculate the different acoustic characteristics. As the corpus is relatively small, we have focused on global characteristics, computed on the whole sentence, rather than local features (sub- and segmental level). We compute the most common ones (Pell et al, 2009): 1) F0 and energy features: mean, minimum, maximum, range, 2) Duration: Articulation rate. To evaluate the vocal characteristics, jitter and shimmer are commonly taken into account as features of the automatic speech recognition system. Jitter (respect. Shimmer) measures the perturbation of the length (respect. amplitude) of two consecutive pitch periods. In other words, jitter represents the varying pitch in the voice and Shimmer stands for the varying loudness in the voice. These features are computed for each sentence, and the mean value is used for each emotion.

Figure 11 and Figure 12 (detailed in the table 3) show a summary representation of F0 for each emotion. The small 95% confidence interval shows that F0 mean is quite stable for all sentences, thus this is a robust characteristic. Moreover, a cross-species pattern was found by the ethologist Eugene Morton, in which high

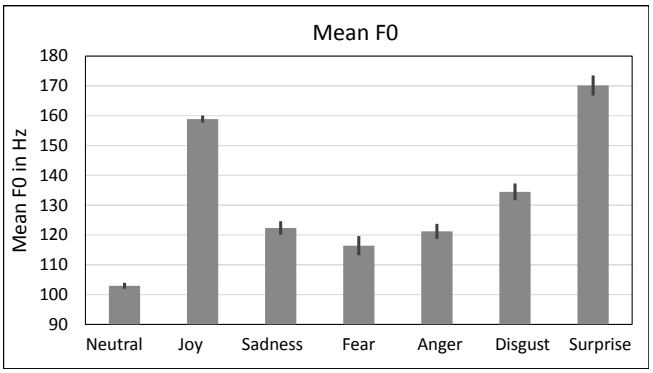

Fig. 11 Mean F0 values of the 7 emotions and their 95% confidence interval.

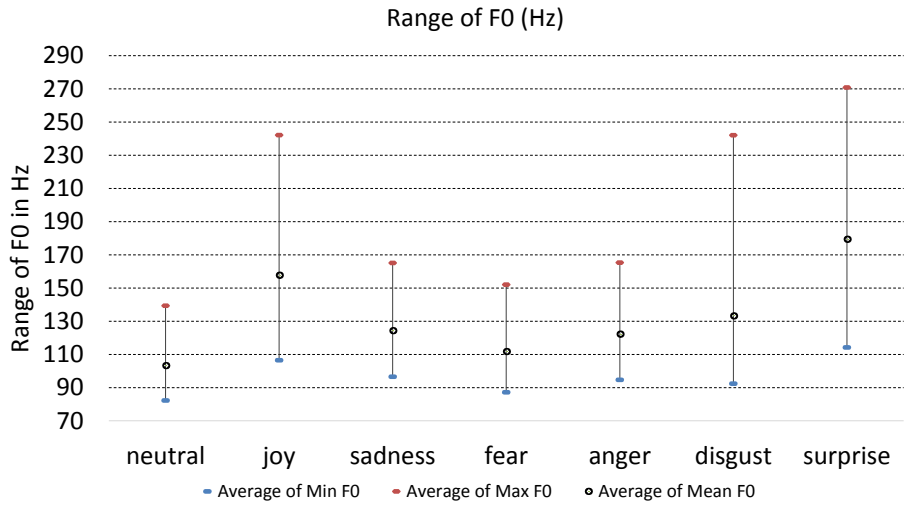

Fig. 12 Range F0 values of the 7 emotions.

	neutral	**joy**	**sadness**	**fear**	**anger**	**disgust**	**surprise**
Average Min F0	82.19	106.49	96.53	87.16	94.68	92.37	114.18
Average Max F0	139.33	242.08	165.07	152.01	165.31	242.03	270.77
Average Mean F0	103.20	157.71	124.24	111.731	122.21	133.14	179.37

Table 3 Range F0 values of the 7 emotions.

frequencies are correlated with affiliative behaviors, whereas low frequencies are associated with aggressive behaviors (Morton, 1977, 1994). Also, in a cross-cultural study of speech prosody, Bolinger (1978) found a similar association in humans: A high F0 is associated with friendly behaviors whereas a low F0 is associated with aggressive behaviors. The findings plotted in figures 11 and 12 reveal that joy and surprise have greater F0 than the other emotions, as suggested previously. As for fear and anger, they have the lowest F0 since they are the farthest from pleasant attitude. Nevertheless, summarizing intonation of emotion by min/max/mean values is quite restrictive. The intonation can also be considered like a gesture. The contour of intonation is also meaningful to compare emotions. An example of the typical F0 contour of expressive French utterance Mais les gens ne se mettent pas en grève par plaisir, tu devrais le savoir, si les grenouilles avaient des ailes, elles ne s'embêteraient pas à sauter. (But people do not go on strike for pleasure, you should know, if the frogs had wings, they would not bother to jump) are plotted in Figure 13. The mean F0 value has been calculated for each syllable using PRAAT. The duration of syllable is not represented in the figure, i.e., the utterance duration for each emotion is different. The contours of the 10 utterances of all the emotions have been computed. After that, the mean F0 of each utterance has been calculated and plotted in Figure 14. This figure captures the same trend in average than inside a sentence (Figure 13). The ranking of emotions in terms of F0 values is ordered almost in the same way for all sentences. It is noticeable that the dominant tendencies of F0 contours are correctly captured by the mean measure and the global F0 rank of each emotion is preserved.

Globally, the obtained results confirm what has been found in other studies (Scherer, 1986; Paeschke et al, 1999) about the correlation between the emotions

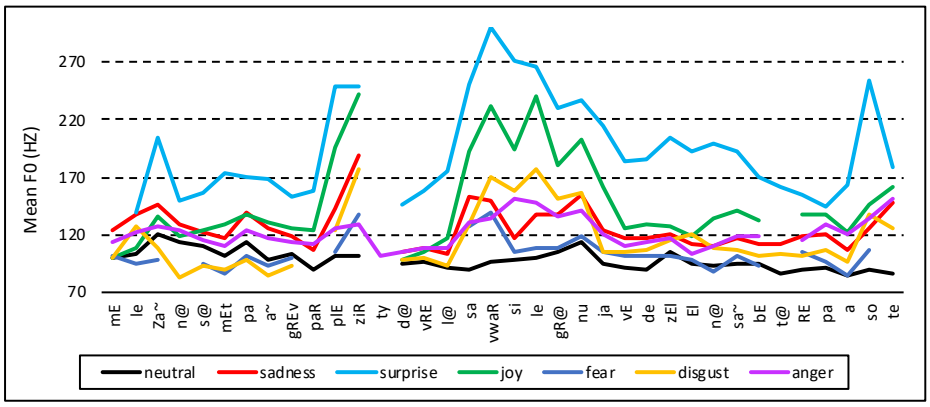

Fig. 13 F0 contours of a corpus sentence for the 7 emotions (per syllables).

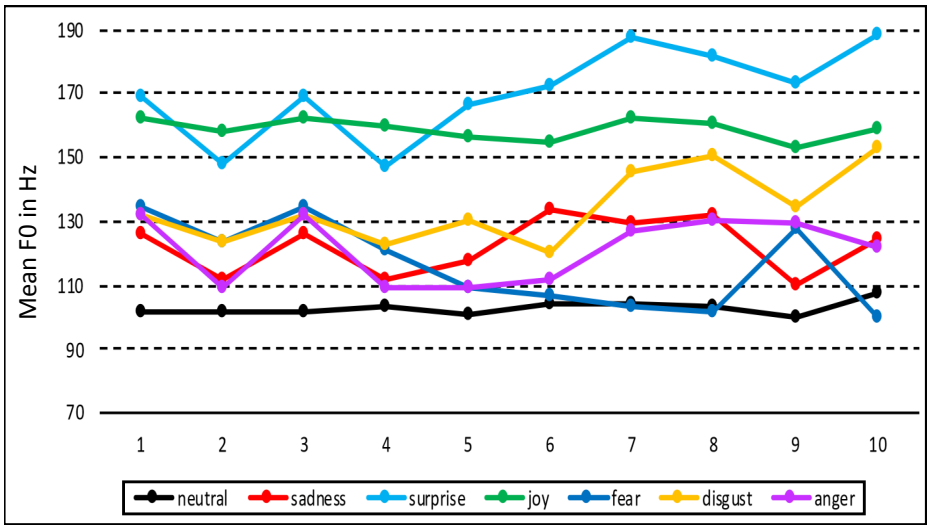

Fig. 14 F0 mean values for the 7 emotions (per sentences)

and the average F0. All the emotions have a higher global F0 mean compared to neutral. This result is similar to what is commonly expected, except for sadness which was deemed to have lower F0 than neutral. This difference may come from the acted (exaggerated) nature of our data corpus that can convey a stronger degree of affect than spontaneous speech/emotion.

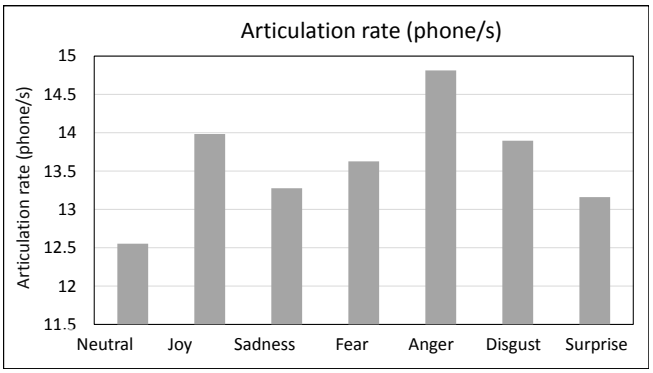

Fig. 15 Articulation rate per emotion (number of sounds per second), calculated from the 10 sentences.

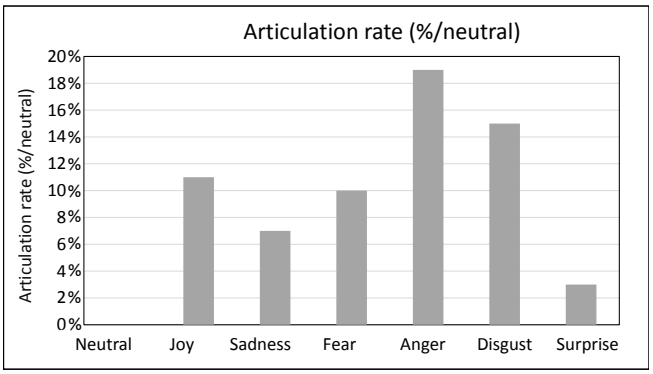

Fig. 16 Articulation rate per emotion according to the neutral articulation rate, calculated from the 10 sentences.

Figure 15 and table 4 show the statistics of the speech rate (phones per second). To calculate this measure, the total length of sounds (silences were not counted) was divided by the total number of sounds per sentence. This information (a length of a single sound in a given emotion) was used to calculate the number of sounds per second. The mean value has been calculated for the ten sentences for all the emotions. All the emotions have greater speech rate than neutral. In Figure 16, articulation rate is expressed in terms of percentage according to the articulation

	anger	**disgust**	**fear**	**joy**	**neutral**	**sadness**	**surprise**
(phone/s)	15.24	14.73	14.1	14.19	12.82	13.66	13.22
(%/neutral)	19%	15%	10%	11%	0%	7%	3%

Table 4 Articulation rate per emotion calculated from the 10 sentences.

rate of neutral emotion. For instance, +20% means that the speed is 20% faster than the neutral speed.

Articulation rate differentiates *anger* and *disgust* from other emotions and are clear faster than *neutral*. *Joy* and *fear* have very similar speed around 10% faster. *Sadness* and *surprise* have lower acceleration (respectively 7% and 3%). However, unlike past studies, in our analysis, *sadness* and disgust were found to be associated with faster, rather than a slower speaking rate compared to *neutral*.

Figures 17 and 18 present the results of our data for jitter and shimmer parameters. The difference between the emotions are very subtle for these features. Results showed that higher jitter and shimmer values are associated with *fear*, *disgust* and *anger*. The lowest jitter and shimmer value was found for *surprise* emotion. For the other emotions (*sadness* and *joy*), they have similar value to *neutral* (*sadness* had a slightly bigger value than *joy*). Furthermore, Nunes (2013) found that jitter and shimmer are higher for the most negative emotions and low for the positive one. Our findings are in line with these results, except for sadness that represents a low jitter and shimmer values (close to *neutral*).

Some emotions are distinguished by their acoustic features while others share similar characteristics:

Neutral: neutral emotion was used again as a reference to compare the behavior of the examined features when a certain emotion is activated. The lowest features

values were found for this emotion (lowest F0, articulation rate, shimmer and jitter).

Surprise: This emotion exhibits the highest F0 value. *Surprise* was also the only one of all emotions to have a lower shimmer and jitter values than *neutral*.

Joy: The F0 values were also important. But, unlike *surprise*, the articulation rate of *joy* was elevated. Shimmer and jitter values were identical to *neutral*.

Sadness: This emotion has a modest F0 mean and articulation rate values. As for shimmer and jitter, *sadness* values were very close to *neutral*.

Disgust: This emotion has a mean F0 value greater than *neutral* but the range of F0 value is clearly one of the highest ones. The articulation rate is also one of the fastest ones.

Anger: when analyzing anger sentences, the F0 range was relatively low, yet the articulation rate was found to be the highest among all the other emotions. Its jitter value is identical to *neutral* but the shimmer value is more important.

Fear: For fear, F0 value was very low, slightly above *neutral*. Unlike *neutral*, F0 peaks for *fear* are sharper and reach higher levels. On the other hand, an average articulation rate value was found for *fear* emotion. This emotion has the highest jitter and shimmer values.

3.2 Perceptual Evaluation

The main purpose of the perceptual evaluation is to assess whether the actor was able to convey the different emotions correctly to the human receivers. This is

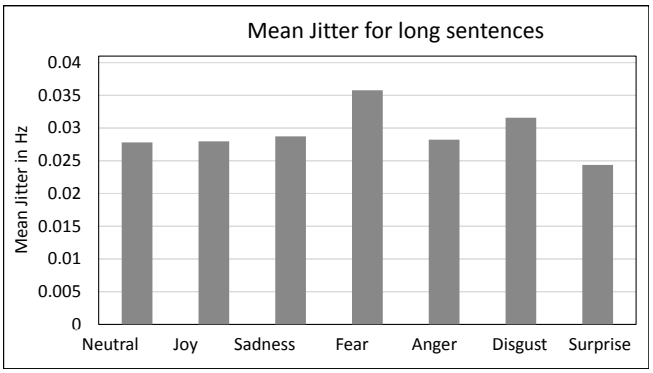

Fig. 17 Mean jitter value for the 7 emotions.

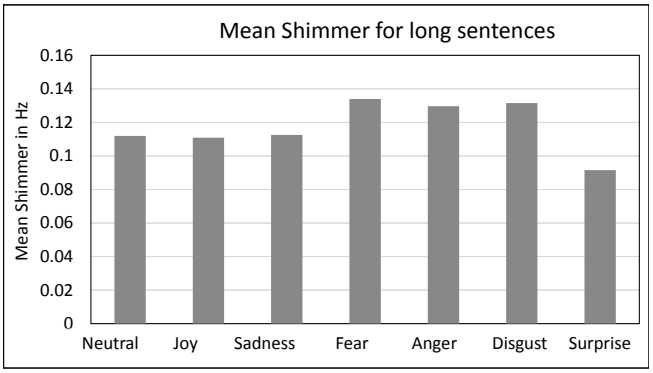

Fig. 18 Mean shimmer value for the 7 emotions.

essential to make a decision about the quality of the acquired expressive audiovisual corpus. As the corpus will be used for expressive audiovisual speech synthesis, it is important to evaluate the configuration of the 3D markers, to see whether this minimal presentation of the face is sufficient to model facial expressivity, and whether the chosen number and positions of sensors are acceptable to capture expressivity features correctly. Similar evaluation technique has been used in the past to evaluate an audiovisual speech synthesis system (Bailly et al, 2002). In a previous work, we have presented a detailed study focusing on the perceptual aspects of expressivity (Ouni et al, 2017). We have also discussed the influence of

each modality and its contribution to the perception of emotions. In particular, we have compared two modalities: bimodal (with audio) and unimodal (without audio), across three presentations: 3D points without head movement, 3D points with head movement and the video of the actor's face. The stimuli that have been used were 10 short sentences for the different emotions (all the details can be found in Ouni et al (2017)). In the current evaluation, we will focus only on two presentations and two modalities: (1) short vs. (2) long utterances, and (3) actor face (video) vs. (4) 3D points representing the face of the actor (a minimal presentation of the face).

3.2.1 Stimuli

Twenty sentences (10 short and 10 long) were uttered by a 27-year-old, semi-professional, male actor with 7 different emotional states (neutral, joy, surprise, fear, anger, sadness, and disgust). The actor has used the technique *exercise in style*. As participants were not all from the same culture, we choose to use the basic emotions (neutral, joy, surprise, fear, anger, sadness and disgust) since they were proven to be universal (Ekman and Friesen, 1971). For each emotion, two recordings of ten sentences (short and long) were made. We cut the recording into one sentence per stimuli file, then three types of stimulus were presented to the participants: (1) the real actor's face audiovisual stimulus (2) the minimalist face representation audiovisual stimulus (3) the minimalist face representation visual only stimulus. For each part of the experiment, the stimulus was presented randomly.

3.2.2 Participants

The perceptual experiments were conducted with two groups of participants. The perceptual experiment of long sentences, counted twelve naive participants (adults, 5 females, 7 males). For short sentences another group of thirteen participants (adults, 3 females, 10 males) participated in the experiment. The participants from the two groups were not necessarily native French speakers, but they were living in France during that period of time.

3.2.3 Method

During the perceptual experiments, each participant connected to a web application that we created. A series of stimuli were presented one by one. Participants had to choose from a list of seven emotions (neutral, joy, surprise, fear, anger, sadness and disgust) the one expressed by each stimulus. They had to select an answer and validate it to be able to see the next stimuli. It is possible to play each stimulus as many times as wanted. We did not comment or impose any definitions of the emotions to avoid biasing their perception. The participants passed the tests with audio only, then visual only before passing the audiovisual tests.

3.2.4 Results

After collecting the results from all the participants, we computed the statistical significance levels using the p-values from the t-test and we corrected them using Holm Bonferoni method (Holm, 1979). For each experiment, we used a degree of freedom equal to the number of participants minus one (12 for short sentences and 11 for long ones). We considered an alpha equal to 5% and a chance level

		Perceived emotion						
		joy	**sadness**	**anger**	**fear**	**disgust**	**surprise**	**neutral**
Produced emotion	**joy**	**95.00(*)**	0	0	0	0	5	0
	sadness	0	**83.33(*)**	0	3.33	10	0	3.33
	anger	0	8.33	**60(*)**	1.67	25	3.33	1.67
	fear	0	1.67	8.33	**71.67(*)**	1.67	16.67	0
	disgust	0	0	48.33	0	**51.67(*)**	0	0
	surprise	10	0	0	1.67	0	**86.67(*)**	1.67
	neutral	0	5	0	0	0	0	**95(*)**

Table 5 The confusion matrices of the recognition rate of 7 emotions with **the real actor's face video for long sentences**. The columns represent the distribution of the answers given by the participants.

		Perceived emotion						
		joy	**sadness**	**anger**	**fear**	**disgust**	**surprise**	**neutral**
Produced emotion	**joy**	**97.12(*)**	0	0	0	0	2.88	0
	sadness	0	**83.65(*)**	0	0	14.42	0	1.92
	anger	0	2.88	**70.19(*)**	2.88	9.62	7.69	6.73
	fear	0	2.88	0	**92.31(*)**	0	4.81	0
	disgust	0	4.81	4.81	0.96	**89.42(*)**	0	0
	surprise	4.81	0	15.38	0	4.81	**75(*)**	0
	neutral	0	0	0	0	0.96	0	**99.04(*)**

Table 6 The confusion matrices of the recognition rate of 7 emotions with **the real actor's face video for short sentences**. The columns represent the distribution of the answers given by the participants.

of 14%. We added an (*) symbol to statistically significant recognition rates in Tables 5– 10 and (-) for non-significant recognition rates. For the audiovisual long sentences, Table 5 represents the results of the participants answers regarding audiovisual videos of the real actors face. Table 5 shows that joy, sadness, fear, surprise and neutral were very well recognized. All of these emotions have more

		Perceived emotion						
		joy	**sadness**	**anger**	**fear**	**disgust**	**surprise**	**neutral**
Produced emotion	**joy**	**81.67(*)**	0	0	1.67	0	10	6.67
	sadness	0	**76.67(*)**	0	3.33	15	1.67	3.33
	anger	1.67	5	**46.67(*)**	6.67	20	16.67	3.33
	fear	0	5	5	**66.67(*)**	5	18.33	0
	disgust	0	0	65	6.67	**23.33(-)**	1.67	3.33
	surprise	16.67	1.67	0	5	3.33	**68.33(*)**	5
	neutral	0	16.67	0	1.67	5	0	**76.67(*)**

Table 7 The confusion matrices of the recognition rate of 7 emotions with **the minimalist face representation for long sentences (with audio)**. The columns represent the distribution of the answers given by the participants.

		Perceived emotion						
		joy	**sadness**	**anger**	**fear**	**disgust**	**surprise**	**neutral**
Produced emotion	**joy**	**73.33(*)**	0	0	1.67	0	15	10
	sadness	1.67	**50(*)**	1.67	11.67	23.33	0	11.66
	anger	0	8.33	**18.33(-)**	10	13.33	33.33	16.67
	fear	5	10	5	**30(-)**	6.67	23.33	20
	disgust	0	3.33	70	1.67	**21.67(-)**	0	3.33
	surprise	15	3.33	6.67	16.67	1.67	**38.33(*)**	18.33
	neutral	0	25	1.67	10	11.67	0	**51.67(*)**

Table 8 The confusion matrices of the recognition rate of 7 emotions with **the minimalist face representation for long sentences without audio**. The columns represent the distribution of the answers given by the participants.

than 70% recognition rate on the diagonal value of the confusion matrix. As for anger and disgust, those emotions were more confused, which is consistent with the literature (Ekman and Friesen, 1986). Some research revealed that it is incorrect to assume that public share comparable meanings of the emotion terms used, and

		Perceived emotion						
		joy	**sadness**	**anger**	**fear**	**disgust**	**surprise**	**neutral**
Produced emotion	**joy**	**84.62(*)**	0	0	0.96	1.92	11.54	0.96
	sadness	0	**58.65(*)**	0.96	0	25.96	0	14.42
	anger	0.96	1.92	**40.38(*)**	10.58	18.27	7.69	20.19
	fear	0.96	4.81	0	**75.00(*)**	4.81	13.46	0.96
	disgust	0.96	2.88	39.42	1.92	**47.12(*)**	1.92	5.77
	surprise	3.85	0	3.85	9.62	3.85	**77.88(*)**	0.96
	neutral	0	3.85	2.88	0	4.81	0	**88.46(*)**

Table 9 The confusion matrices of the recognition rate of 7 emotions with **the minimalist face representation for short sentences (with audio)**. The columns represent the distribution of the answers given by the participants.

		Perceived emotion						
		joy	**sadness**	**anger**	**fear**	**disgust**	**surprise**	**neutral**
Produced emotion	**joy**	**75(*)**	0.96	0.96	2.88	0.96	15.38	3.85
	sadness	0.96	**69.23(*)**	0	4.81	14.42	0	10.58
	anger	2.88	13.46	**5.77(-)**	16.35	4.81	47.12	9.62
	fear	2.88	8.65	7.69	**21.15(-)**	14.42	7.69	37.5
	disgust	0	4.81	72.12	0.96	**17.31(-)**	2.88	1.92
	surprise	5.77	4.81	4.81	34.62	5.77	**40.38(*)**	3.85
	neutral	1.92	9.62	8.65	5.77	11.54	0	**62.5(*)**

Table 10 The confusion matrices of the recognition rate of 7 emotions with **the minimalist face representation for short sentences without audio**. The columns represent the distribution of the answers given by the participants.

that the common understanding of the word disgust reflects a combination of the conceptual meanings of disgust and anger (Nabi, 2002).

As can be seen in Table 6, audiovisual short sentence presentation has higher recognition rates for almost all the emotions. This can be explained by the fact that

the actor produced more intense emotion, as the duration to deliver the emotions is short. Regarding the audiovisual corpus, the perceptual experiment results show that the majority of participants validate the performance of the actor. It confirms the good quality of the produced audiovisual speech. Interestingly enough, the meaning of the sentences did not have a visible impact on the emotion recognition rate which confirms that the exercise-in-style technique can be used to record a corpus for speech synthesis with a unique textual corpus for the six emotions and the neutral state. We recall that when developing the expressive talking head, we will not use the video of the actor directly, but the 3D points corresponding to the markers and sensors on the face of the actor (Figure 2) represented by the layout of Figure 2. The 3D points will be used to animate a 3D model using the animation technique presented by Ouni and Gris (2018), for instance. The purpose of this perceptual evaluation is also to see whether the landmarks are sufficient to express the different emotions. Although this information is minimalist, it can give a preview of the quality of the talking head rendering, as it will be based on this data. We included an evaluation of this minimal representation in the perceptual evaluation. The data is processed as explained in section 2.1.

For long sentences, all the emotions were correctly recognized except for disgust (see Table 7). The lower recognition rate were that of anger and disgust that were confused with each other (46% and 23% recognition rate respectively). An average of 15% of recognition rates for the neutral state and the six emotions was lost when the minimal representation was presented rather than the real actor's face. This decline in the recognition rate can be explained by the absence of actual facial muscles in the stimuli, given that the minimalist face representation contains much less information than the original face.

For short sentences, the results presented in Table 9 show that sadness and anger have lower recognition rate than in long sentences. However, joy, fear, disgust, surprise and neutral have higher recognition rate than in long sentence presentation. Anger and sadness, in their minimalist presentation, may need some time to be identified by listeners. Nevertheless, the recognition rate follows globally the same trend for the short and long sentences, but it is difficult to confirm that the sentence length is a factor in emotion perception.

For the experiments using audiovisual minimalist face representation, it may not be clear if the recognition rate observed is mainly driven by the information contained in the voice, or by the minimalist face features. To clarify this, we have conducted other tests with the minimalist face representation without audio. The results are presented in Table 8 and Table 10. The difference in recognition rates between long and short sentences seems to be less obvious. Also, the recognition rate dropped drastically for anger, disgust and fear, in long and short sentences and became statistically insignificant. This shows that for these emotions, the acoustic modality carries most of the emotional information.

These results show that the minimal face representation carries important emotional information for most of the studied emotions. This validates the quality of the minimal representation as the recognition rate reflects the same trend of the full information conveyed by the real actors face recording even with a lower recognition rate.

4 Discussion

In this paper, we have presented a multimodal acquisition technique to collect audiovisual data (motion capture data) to develop a talking head. Some technical limitations of the acquisition systems make it very difficult to record high quality corpus with a single acquisition technique, especially for the (related to speech) articulator's area. We have combined different systems to get the relevant information for each part of the face (EMA for speech-related lip movement, RealSense for eyes and VICON cameras for facial expressions).

As acquiring multimodal data is time-consuming, some considerations are important to take into account before recording a large corpus. In this paper, we share our experience about recording and analyzing a small audiovisual corpus with a single speaker. In the context of expressive audiovisual speech, it is important to assess the acting of the recorded speaker to see whether the emotion expressed by the actor are actually well perceived. In addition, the acquired visual data is a set of 3D points, and we need to assess whether they hold enough information about the different emotional contexts. To do so, we presented a perceptual evaluation preceded by a visual and acoustic analyses of the corpus.

In the speech production analysis, we analyzed the visual and the acoustic features of our data. For the visual modality, we conducted PCA to extract the dominant facial movements of each emotion during speech sequences. We described the first three principal component in terms of FACS-AUs. We found that our conclusions are similar to what is commonly expected. In addition to confirming previous findings, the present study demonstrated that the main facial characteristics of these emotions are maintained even during the speech activity. Moreover,

additional Action Units are present in the first three principal components of all the emotions, and they are related to speech activity (mouth opening, protrusion).

The perceptual experiment results show that the majority of the participants validate the performance of the actor, this confirms the good quality of the produced audiovisual speech. Interestingly enough, the meaning of the sentences did not have a visible impact on the recognition rate of the emotions. This confirms that the exercise-in-style technique can be used to record a corpus for speech synthesis with a unique textual corpus for different emotions. We also found that it is necessary that the actor and the human perceivers share the same definition of emotions (i.e. disgust) to avoid possible confusion based on misunderstanding of emotion names significance.

The minimalist face representation by few 3D points seems to be sufficient to convey reasonably correct emotions. Yet, some of them were mainly carried by the acoustic modality. It is impressive that this minimalist representation was able to express some emotions correctly with high recognition rate. Obviously, it is not possible to reach the performances of the real actors face presentation, as several additional information are missing in this minimalist presentation (muscles, wrinkles, head movement, higher facial coverage, etc.). The landmark layout presented in Figure 2 seems to be well adapted to speech and to facial expressions.

The length of the sentences has some effect on the emotion expression but the recognition rate does not seem to be heavily affected by duration. Based on the results presented in Tables 5, 6, 7 and 9, the expression of emotions is more intense during short utterances, and probably diluted for longer ones, as the actor cannot maintain the same intensity all long. However, it is difficult to confirm that the sentence length is a factor in emotion perception. This issue may need a

dedicated study. Each one of the examined emotions has a specific configuration of the investigated parameters. Some features express certain emotions, such as mouth and eyes opening for visual modality. Audio or visual parameters characterize some emotions more than others and each emotion should be represented with audio and visual features jointly. Some previous studies brought to light the relationship between auditory and visual cues, especially for signaling certain aspects of communication. For instance, Huron and Shanahan (2013) exposed that the prominence of a word is enhanced if a pitch accent is additionally marked with a visual eyebrow movement (Cave et al, 1996). There have been also claims that head movement and eyebrow movements are correlated with acoustic features of prosody, such as fundamental frequency and amplitude (Cave et al, 1996; Yehia et al, 2002).

To conclude, this paper presented a technical description of an audiovisual acquisition system combining three different acquisition techniques. We present a set of analysis and evaluations to assess the acting quality, expressiveness and the spatial representation of the corpus. We believe that sharing our experience and all the details of the acquisition and analysis of an expressive audiovisual corpus can be useful to other researchers wishing to undertake a similar study, particularly in the field of speech synthesis.

Acknowledgements This work was supported by Region Lorraine (COREXP Project), Inria (ADT Plavis) and the EQUIPEX Ortolang. We also thank our actor F.S. for his participation in this study.

References

Bailly G, Gibert G, Odisio M (2002) Evaluation of movement generation systems using the point-light technique. In: Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on, IEEE, pp 27–30

Bandini A, Ouni S, Cosi P, Orlandi S, Manfredi C (2015) Accuracy of a markerless acquisition technique for studying speech articulators. In Interspeech 2015. In: Interspeech 2015

Barbulescu A (2015) Generation of audio-visual prosody for expressive virtual actors. Theses, Université Grenoble Alpes

Barra Chicote R, Montero Martínez JM, et al (2008) Spanish expressive voices: corpus for emotion research in spanish. In: Second international workshop on emotion: corpora for research on emotion and affect, international conference on language resources and evaluation (LREC 2008)

Berry JJ (2011) Accuracy of the ndi wave speech research system. Journal of Speech, Language, and Hearing Research 54(5):1295–1301

Boersma P, et al (2002) Praat, a system for doing phonetics by computer. Glot international 5

Bolinger D (1978) Intonation across languages. Universals of human language

Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation 42(4):335

Cave C, Guaitella I, Bertrand R, Santi S, Harlay F, Espesser R (1996) About the relationship between eyebrow movements and fo variations. In: Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on

Czyzewski A, Kostek B, Bratoszewski P, Kotus J, Szykulski M (2017) An audio-visual corpus for multimodal automatic speech recognition. Journal of Intelligent Information Systems 49(2):167–192

Dutoit T (2008) Corpus-based speech synthesis. In: Springer Handbook of Speech Processing, Springer, pp 437–456

Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. Journal of personality and social psychology 17(2):124

Ekman P, Friesen WV (1976) Measuring facial movement. Environmental psychology and nonverbal behavior 1(1):56–75

Ekman P, Friesen WV (1986) A new pan-cultural facial expression of emotion. Motivation and emotion 10(2):159–168

Ekman P, Friesen W, Hager J (2002) Facial action coding system: Research nexus. Network Research Information, Salt Lake City, UT 1

Feng Y, Max L (2014) Accuracy and precision of a custom camera-based system for 2-d and 3-d motion tracking during speech and nonspeech motor tasks. Journal of Speech, Language, and Hearing Research 57(2):426–438

Fernandez-Lopez A, Sukno FM (2018) Survey on automatic lip-reading in the era of deep learning. Image and Vision Computing

François H, Boëffard O (2001) Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem. In: Seventh European Conference on Speech Communication and Technology

Hess U, Thibault P (2009) Why the same expression may not mean the same when shown on different faces or seen by different people. In: Affective information processing, Springer, pp 145–158

Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics pp 65–70

Huron D, Shanahan D (2013) Eyebrow movements and vocal pitch height: evidence consistent with an ethological signal. The Journal of the Acoustical Society of America 133(5):2947–2952

Jiang J, Alwan A, Keating P, Auer E, Bernstein L (2002) On the relationship between face movements, tongue movements, and speech acoustics. EURASIP Journal on Applied Signal Processing 11:1174–1188

Jonathan Chevelu OB Nelly Barbot, Delhay A (2008) Expressive prosody for unit-selection speech synthesis. In: LREC

Katz W, Campbell TF, Wang J, Farrar E, Eubanks JC, Balasubramanian A, Prabhakaran B, Rennaker R (2014) Opti-speech: A real-time, 3d visual feedback system for speech training. In: Fifteenth Annual Conference of the International Speech Communication Association

Kawaler M, Czyzewski A (2019) Database of speech and facial expressions recorded with optimized face motion capture settings. Journal of Intelligent Information Systems pp 1–24

Keselman L, Iselin Woodfill J, Grunnet-Jepsen A, Bhowmik A (2017) Intel realsense stereoscopic depth cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 1–10

Lamere P, Kwok P, Gouvea E, Raj B, Singh R, Walker W, Warmuth M, Wolf P (2003) The cmu sphinx-4 speech recognition system. In: IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong

Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): A complete dataset for action unit and

emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, IEEE, pp 94–101

Ma J, Cole R, Pellom B, Ward W, Wise B (2006) Accurate visible speech synthesis based on concatenating variable length motion capture data. IEEE Transactions on Visualization and Computer Graphics 12(2):266–276

Mattheyses W, Latacz L, Verhelst W (2009) On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. EURASIP Journal on Audio, Speech, and Music Processing

Mefferd A (2015) Articulatory-to-acoustic relations in talkers with dysarthria: A first analysis. Journal of Speech, Language, and Hearing Research 58(3):576–589

Mehrabian A (2008) Communication without words. Communication theory pp 193–200

Moore S (1984) The Stanislavski system: The professional training of an actor. Penguin

Morton ES (1977) On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. The American Naturalist 111(981):855–869

Morton ES (1994) Sound symbolism and its role in non-human vertebrate communication. Sound symbolism pp 348–365

Nabi RL (2002) The theoretical versus the lay meaning of disgust: Implications for emotion research. Cognition & Emotion 16(5):695–703

Nunes AMB (2013) Cross-linguistic and cultural effects on the perception of emotions. International Journal of Science Commerce and Humanities 1(8):107–120

Ouni S, Dahmani S (2016) Is markerless acquisition technique adequate for speech production? The Journal of the Acoustical Society of America 139(6):234–239

Ouni S, Gris G (2018) Dynamic Lip Animation from a Limited number of Control Points: Towards an Effective Audiovisual Spoken Communication. Speech Communication 96

Ouni S, Colotte V, Musti U, Toutios A, Wrobel-Dautcourt B, Berger MO, Lavecchia C (2013) Acoustic-visual synthesis technique using bimodal unit-selection. EURASIP Journal on Audio, Speech, and Music Processing

Ouni S, Colotte V, Dahmani S, Azzi S (2016) Acoustic and Visual Analysis of Expressive Speech: A Case Study of French Acted Speech. In: Interspeech 2016

Ouni S, Dahmani S, Colotte V (2017) On the quality of an expressive audiovisual corpus: a case study of acted speech. In: The 14th International Conference on Auditory-Visual Speech Processing

Paeschke A, Kienast M, Sendlmeier WF, et al (1999) F0-contours in emotional speech. In: Proc. 14th Int. Congress of Phonetic Sciences, vol 2, pp 929–932

Pell MD, Paulmann S, Dara C, Alasseri A (2009) Factors in the recognition of vocally expressed emotions: A comparison of four languages. Journal of Phonetics

Queneau R (2018) Exercises in style. Alma Books

Raymond Q (1947) Exercices de style

Schabus D, Pucher M (2014) Joint audiovisual hidden semi-markov model-based speech synthesis. IEEE Journal of Selected Topics in Signal Processing

Scherer KR (1986) Vocal affect expression: A review and a model for future research. Psychological bulletin 99(2):143

Stella M, Stella A, Sigona F, Bernardini P, Grimaldi M, Fivela BG (2013) Electromagnetic articulography with ag500 and ag501. In: Interspeech, pp 1316–1320

Tian YI, Kanade T, Cohn JF (2001) Recognizing action units for facial expression analysis. IEEE Transactions on pattern analysis and machine intelligence

Vatikiotis-Bateson E, Munhall K, Ostry D (1993) Optoelectronic measurement of orofacial motions during speech production. The Journal of the Acoustical Society of America 93(4):2414–2414

Volker Strom RC, King S (2006) Expressive prosody for unit-selection speech synthesis. In: INTERSPEECH

Walsh B, Smith A (2012) Basic parameters of articulatory movements and acoustics in individuals with parkinson's disease. Movement Disorders 27(7):843–850

Wiggers M (1982) Judgments of facial expressions of emotion predicted from facial behavior. Journal of Nonverbal Behavior 7(2):101–116

Yehia HC, Kuratate T, Vatikiotis-Bateson E (2002) Linking facial animation, head motion and speech acoustics. Journal of phonetics 30(3):555–568

Yunusova Y, Green JR, Mefferd A (2009) Accuracy assessment for ag500, electromagnetic articulograph. Journal of Speech, Language, and Hearing Research